E-Values Expand the Scope of Conformal Prediction

Etienne Gauthier, Francis Bach, Michael I. Jordan

(INRIA)

https://arxiv.org/abs/2503.13050

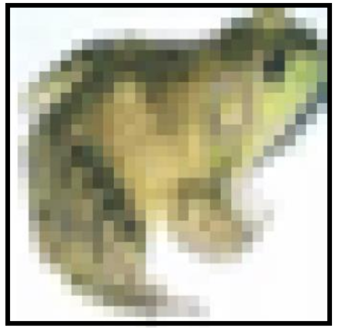
Overview

- Basics of Conformal Prediction
- Batch Anytime-valid Conformal Prediction
- Conformal Prediction with Adaptive Coverage
- Conformal Prediction under Ambiguous Ground Truth

Overview

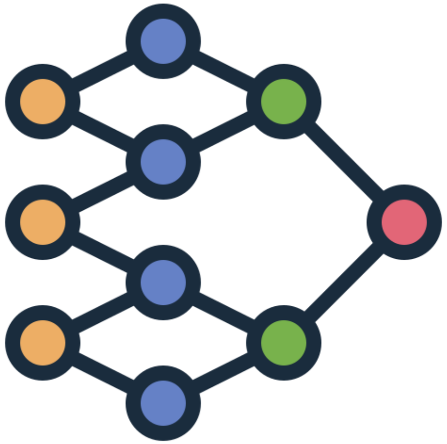
- Basics of Conformal Prediction
- Batch Anytime-valid Conformal Prediction
- Conformal Prediction with Adaptive Coverage
- Conformal Prediction under Ambiguous Ground Truth

Conformal Prediction
Motivation



Test input X_{test}

(Ground truth label Y_{test} = frog)



Predictor f

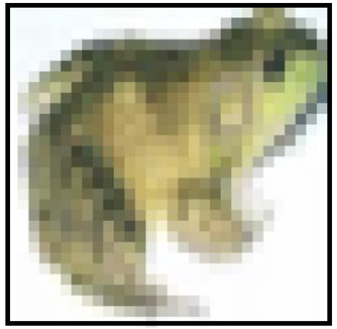
{ airplane, automobile, bird, cat, deer, dog, frog ← Y_{test}, horse, ship, truck }

Prediction

$f\,(X_{test}) = \hat{Y}_{test}$

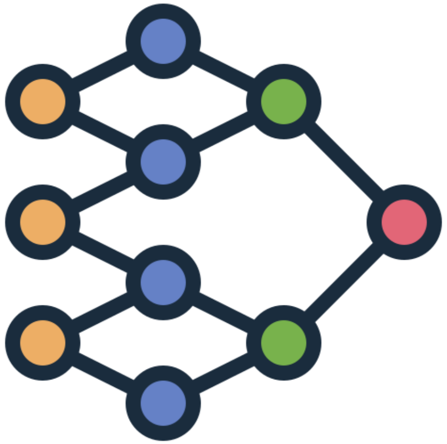
Inria

Conformal Prediction
Motivation



Test input X_{test}

(Ground truth label $Y_{test} =$ frog)



Predictor f

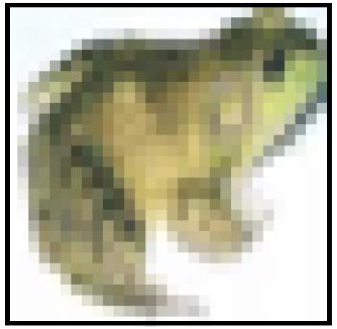
{ airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck } ← Y_{test}

Prediction

$f\,(X_{test}) = \hat{Y}_{test}$

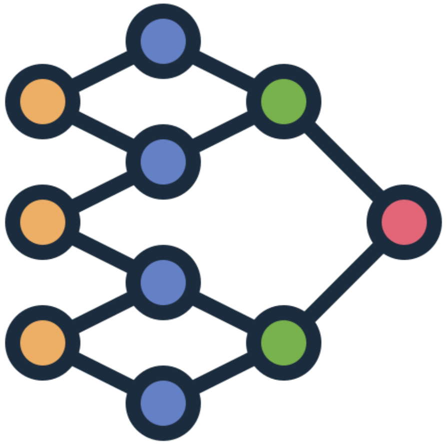
Inria

Conformal Prediction
Motivation



Test input X_{test}

(Ground truth label Y_{test} = frog)



Predictor f

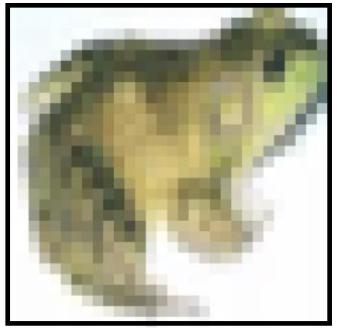
$$\left\{ \begin{array}{c} \text{airplane} \\ \text{automobile} \\ \text{bird} \\ \text{cat} \\ \text{deer} \\ \text{dog} \\ \text{frog} \\ \text{horse} \\ \text{ship} \\ \text{truck} \end{array} \right\}$$

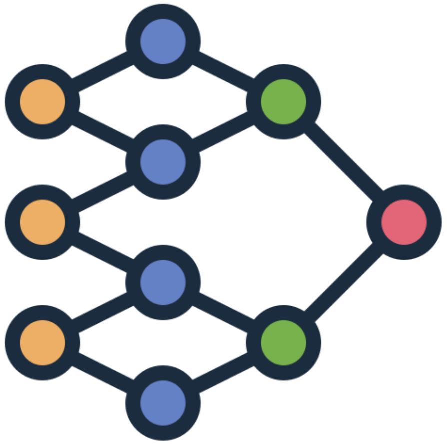
frog ← Y_{test}

Prediction

$f(X_{test}) = \hat{Y}_{test}$

Conformal Prediction
Motivation





$\{$ airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck $\}$

frog $\leftarrow Y_{test}$

➢ **Goal:** build a prediction set $C(X_{test})$ that contains Y_{test} with high probability:
$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$

Te

(Ground truth label Y_{test} = frog)

Predictor f

Prediction

$f(X_{test}) = \hat{Y}_{test}$

Main idea

Symmetric residuals: scores and calibration set

- **Score function $\boldsymbol{S}: \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+^*$**
  - Measures how well the predicted label aligns with the true label
  - Ex: $S(x, y) = -\log p_f(y|x)$ in classification

Main idea

Symmetric residuals: scores and calibration set

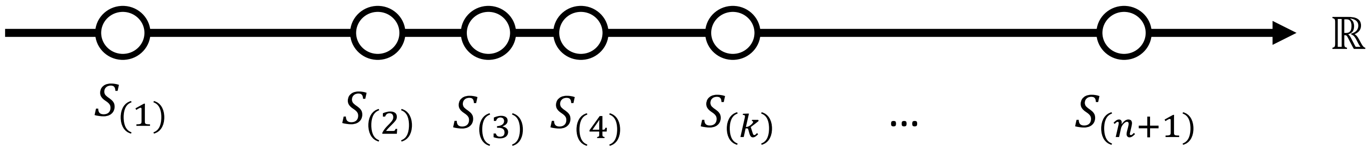
- **Score function $\boldsymbol{S}: \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+^*$**
  - Measures how well the predicted label aligns with the true label
  - Ex: $S(x, y) = -\log p_f(y|x)$ in classification
- **Calibration set**
  - Held-out portion of labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ used to compute prediction sets

Main idea

Symmetric residuals: scores and calibration set

- **Score function $\boldsymbol{S}: \boldsymbol{X} \times \boldsymbol{Y} \rightarrow \mathbb{R}_{+}^{*}$**
  - Measures how well the predicted label aligns with the true label
  - Ex: $S(x, y) = -\log p_f(y|x)$ in classification
- **Calibration set**
  - Held-out portion of labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ used to compute prediction sets
- **Observation:** the $S(X_i, Y_i)$ and $S(X_{test}, Y_{test})$ are i.i.d:

$$\mathbb{P}\big(\text{rank}\big(S(X_{test}, Y_{test})\big) \leq k\big) = \frac{k}{n+1} \;\rightarrow\; k = \lceil (1-\alpha)(n+1) \rceil$$



Main idea
Symmetric residuals: scores and calibration set

- **Score function S: $X \times Y \to \mathbb{R}_+^*$**
  - Measures how well the predicted label aligns with the true label
  - Ex: $S(x, y) = -\log p_f(y|x)$ in classification

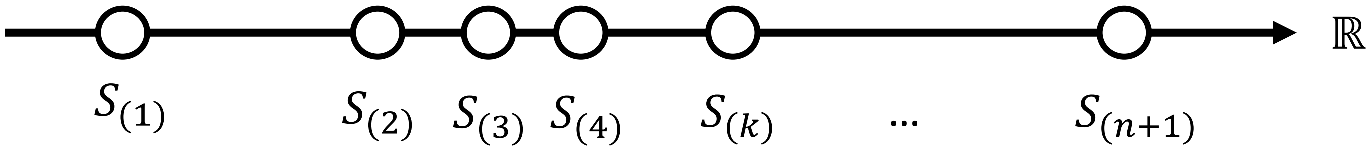
➢ **Conformal Prediction (Vovk et al., 2005):**

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha,$$

where $C(X_{test}) = \{y : \text{rank}(S(X_{test}, y)) \leq \lceil (1-\alpha)(n+1) \rceil\}$.

- **Observation:** the $S(X_i, Y_i)$ and $S(X_{test}, Y_{test})$ are i.i.d:

$$\mathbb{P}(\text{rank}(S(X_{test}, Y_{test})) \leq k) = \frac{k}{n+1} \quad \to \quad k = \lceil (1-\alpha)(n+1) \rceil$$



P-values!

Alternative formulation:

$$\mathbb{P}\left(\frac{1+\sum_{i=1}^{n}\mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}}{n+1}\leq\alpha\right)\leq\alpha$$

P-values!

Alternative formulation:

$$\mathbb{P}\left(\frac{1+\sum_{i=1}^{n} \mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}}{n+1} \leq \alpha\right) \leq \alpha$$

p-value

P-values! And e-values…

Alternative formulation:

$$\mathbb{P}\left(\frac{1+\sum_{i=1}^{n}\mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}}{n+1}\leq\alpha\right)\leq\alpha$$

p-value

$$\mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}=\mathbb{1}\left\{\frac{S(X_i,Y_i)}{S(X_{test},Y_{test})}>1\right\}\leq\frac{S(X_i,Y_i)}{S(X_{test},Y_{test})}$$

$\leq 1/E$ where E is the soft-rank e-value [Wang & Ramdas 2020, Koning 2023, Balinsky & Balinsky 2024]:

$$E=\frac{S(X_{test},Y_{test})}{\frac{1}{n+1}\left(\sum_{i=1}^{n}S(X_i,Y_i)+S(X_{test},Y_{test})\right)}$$

P-values! And e-values…

Alternative formulation:

$$\mathbb{P}\left(\frac{1+\sum_{i=1}^{n}\mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}}{n+1}\leq\alpha\right)\leq\alpha$$

p-value

$$\mathbb{1}\{S(X_i,Y_i)>S(X_{test},Y_{test})\}=\mathbb{1}\left\{\frac{S(X_i,Y_i)}{S(X_{test},Y_{test})}>1\right\}\leq\frac{S(X_i,Y_i)}{S(X_{test},Y_{test})}$$

$\leq 1/E$ where E is the soft-rank e-value [Wang & Ramdas 2020, Koning 2023, Balinsky & Balinsky 2024]:

$$E=\frac{S(X_{test},Y_{test})}{\frac{1}{n+1}\left(\sum_{i=1}^{n}S(X_i,Y_i)+S(X_{test},Y_{test})\right)}$$

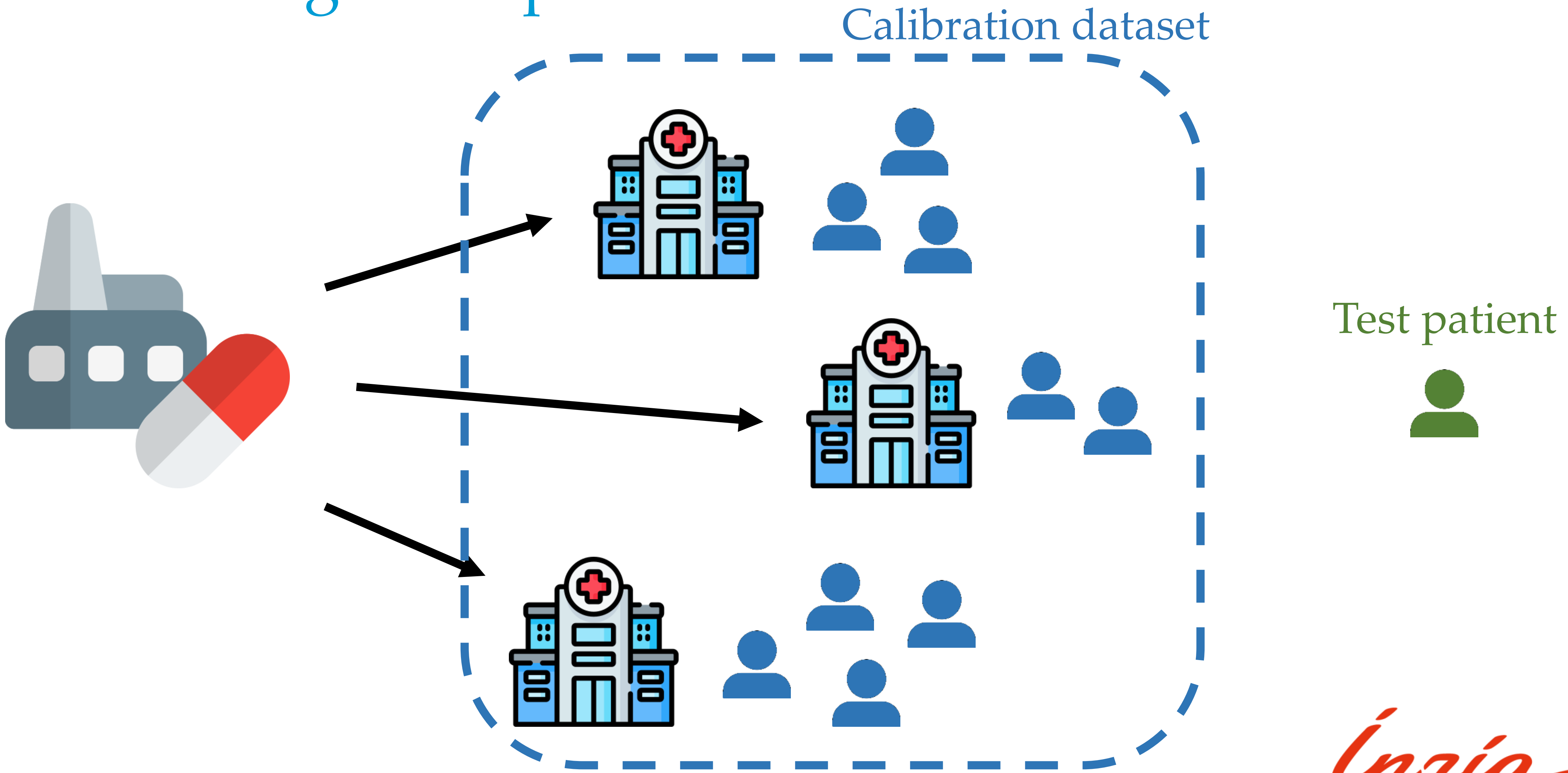
➢ Conformal e-prediction [Vovk 2024]

Inria

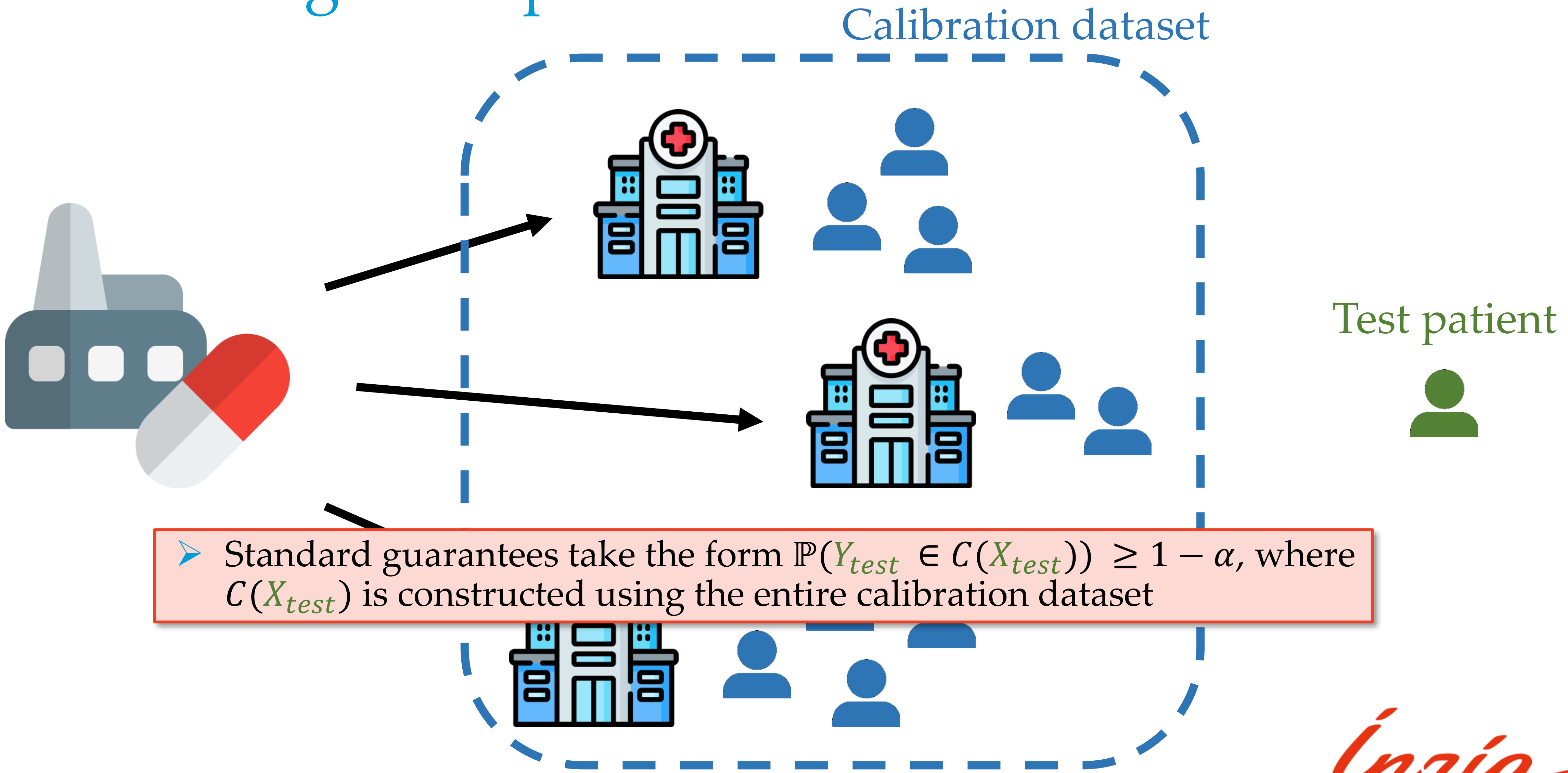
Overview

- Basics of Conformal Prediction
- Batch Anytime-valid Conformal Prediction
- Conformal Prediction with Adaptive Coverage
- Conformal Prediction under Ambiguous Ground Truth

Motivating Example

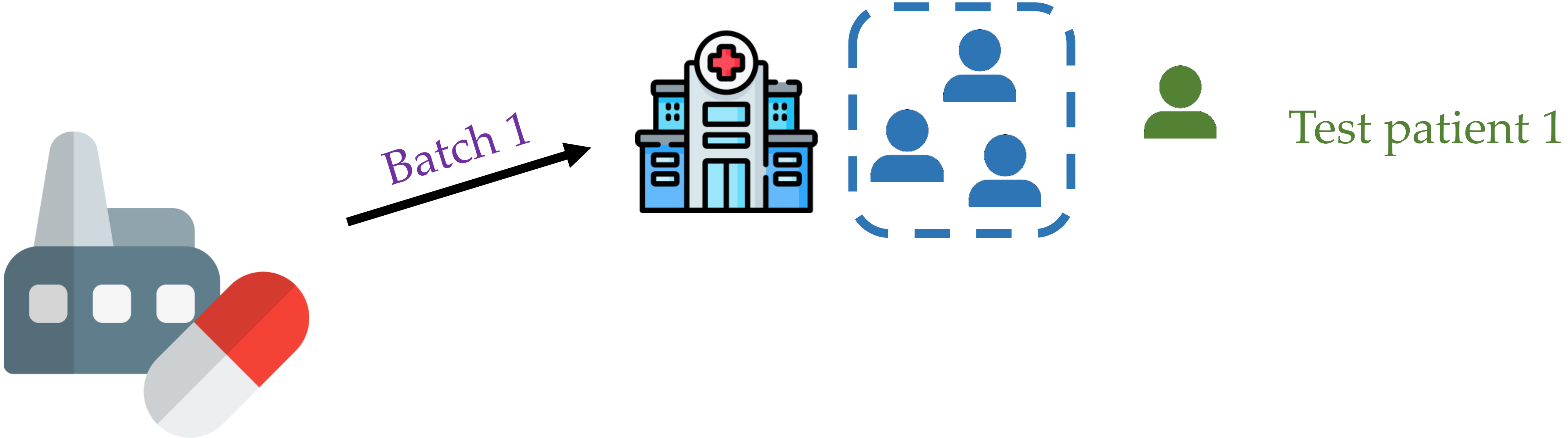


Motivating Example

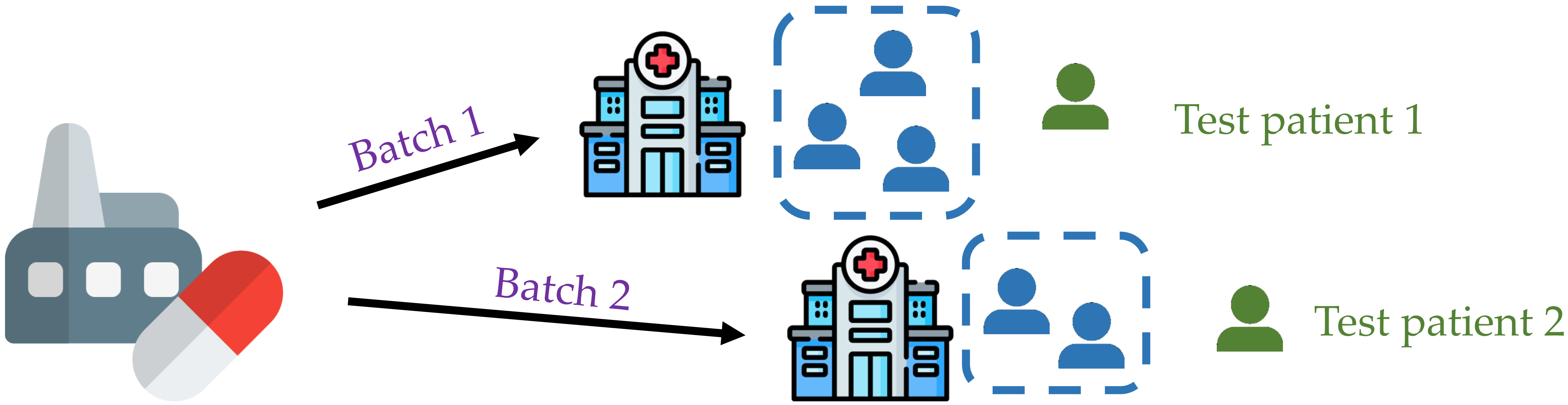


- Standard guarantees take the form $\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$, where $C(X_{test})$ is constructed using the entire calibration dataset

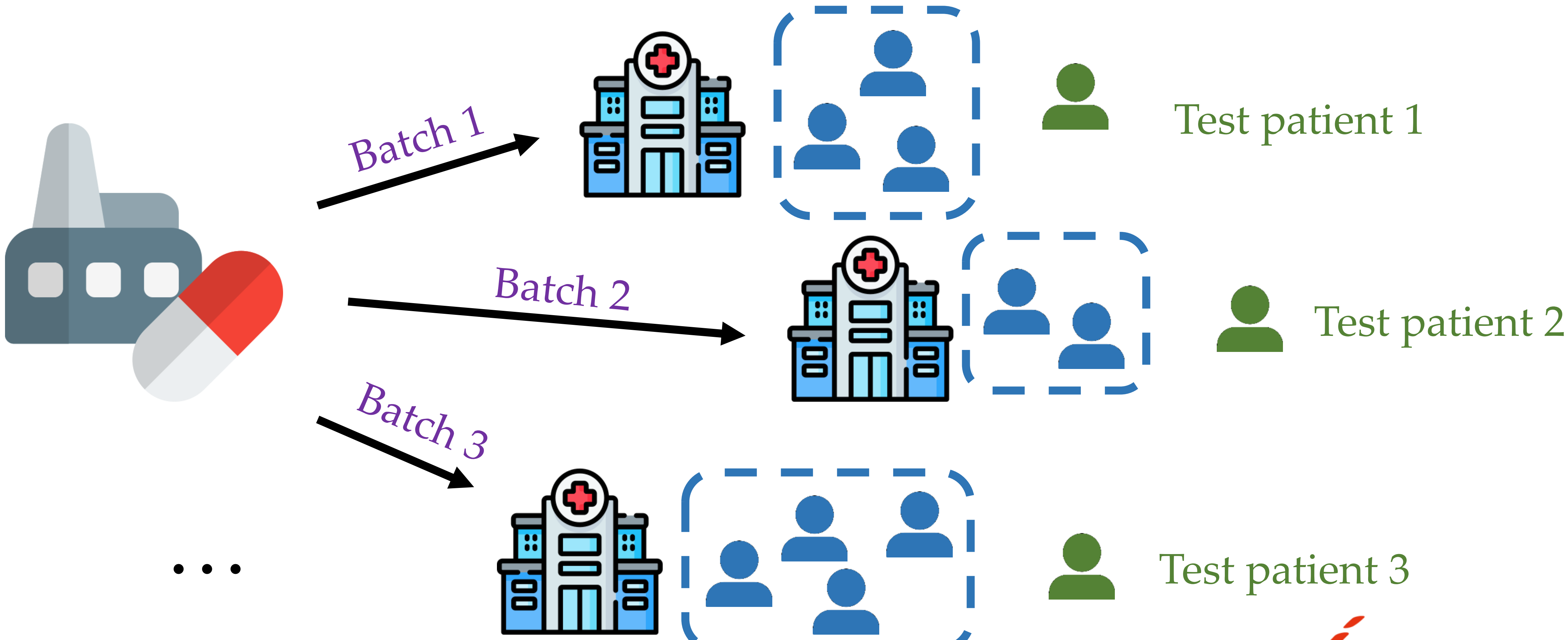
Motivating Example



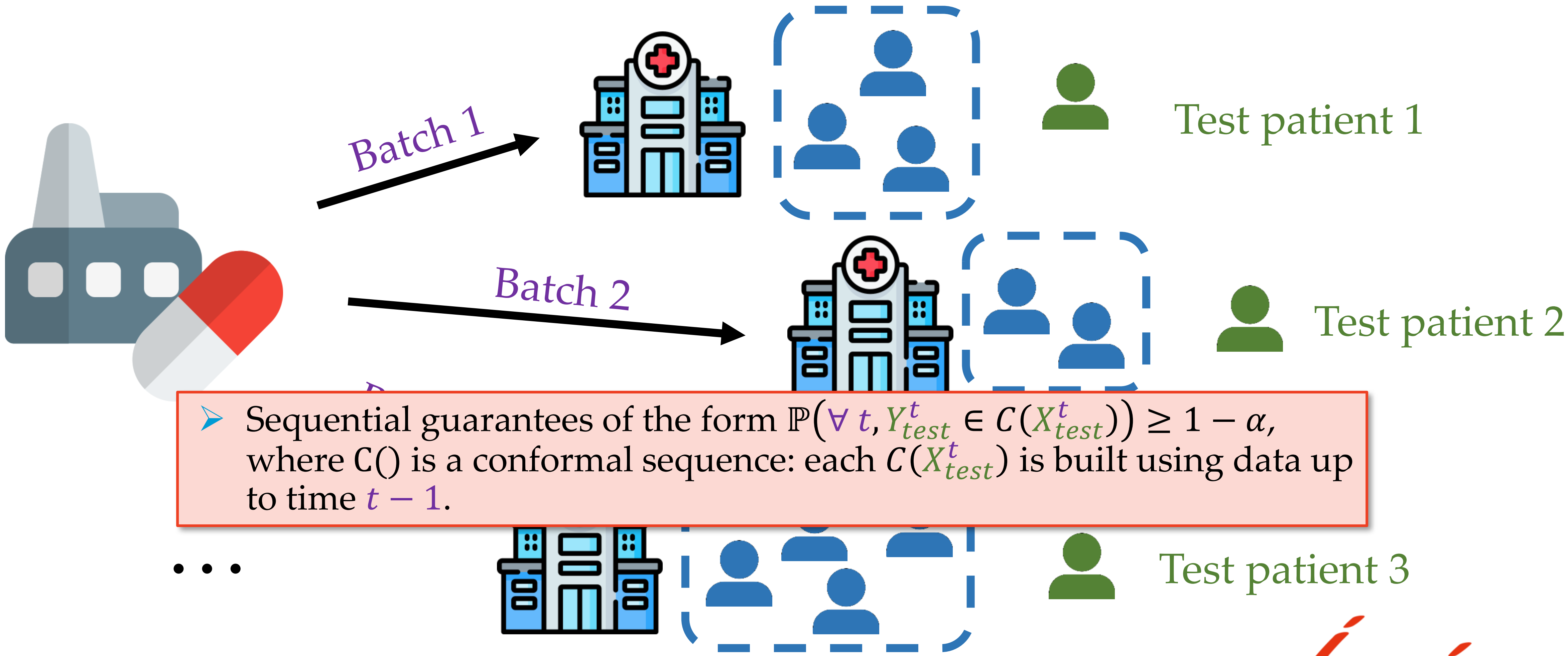
Motivating Example



Motivating Example



Motivating Example



- Sequential guarantees of the form $\mathbb{P}\left(\forall\, t, Y_{test}^{t} \in C(X_{test}^{t})\right) \geq 1-\alpha$, where C() is a conformal sequence: each $C(X_{test}^{t})$ is built using data up to time $t-1$.

Main Result

Supermartingale + Ville's inequality

$$M_t = \prod_{b=1}^{t} (1 - \lambda_b + \lambda_b E_b),$$

$$E_b = \frac{S(X_{test}^b, Y_{test}^b)}{\frac{1}{n_b + 1}\left(\sum_{i=1}^{n_b} S(X_i^b, Y_i^b) + S(X_{test}^b, Y_{test}^b)\right)}$$

Main Result

Supermartingale + Ville's inequality

$$M_t = \prod_{b=1}^{t} (1 - \lambda_b + \lambda_b E_b),$$

$$E_b = \frac{S(X_{test}^b, Y_{test}^b)}{\frac{1}{n_b + 1}\left(\sum_{i=1}^{n_b} S(X_i^b, Y_i^b) + S(X_{test}^b, Y_{test}^b)\right)}$$

➢ **Batch Anytime-valid Conformal Prediction:**

$$\mathbb{P}\left(\forall\, t, Y_{test}^t \in C(X_{test}^t)\right) \geq 1 - \alpha,$$

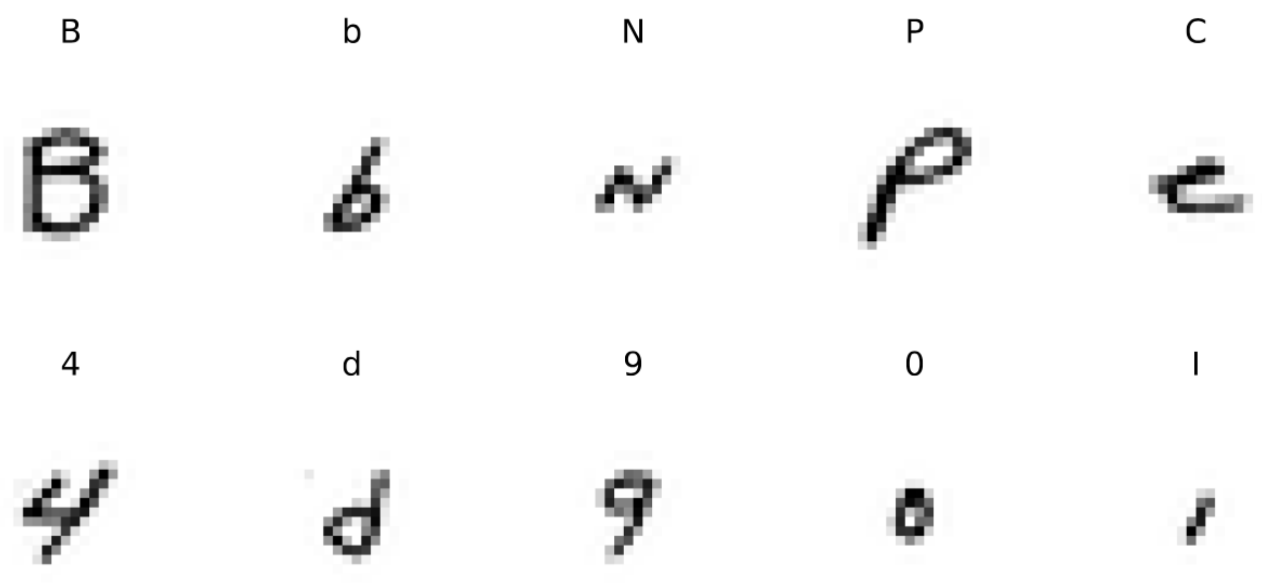
where $C(X_{test}^t) = \left\{ y : \prod_{b=1}^{t-1}(1 - \lambda_b + \lambda_b E_b) \times \frac{S(X_{test}^t, y)}{\frac{1}{n_t+1}\left(\sum_{i=1}^{n_t} S(X_i^t, Y_i^t) + S(X_{test}^t, y)\right)} < 1/\alpha \right\}.$

Ville's inequality

Inria

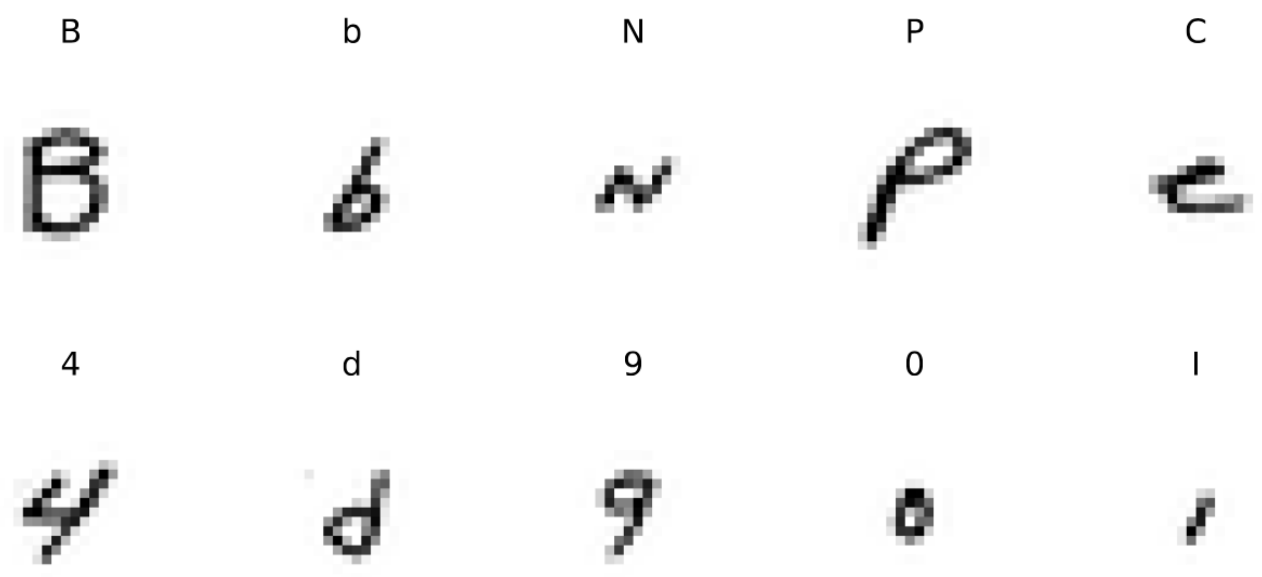
Experiments

- FEMNIST dataset



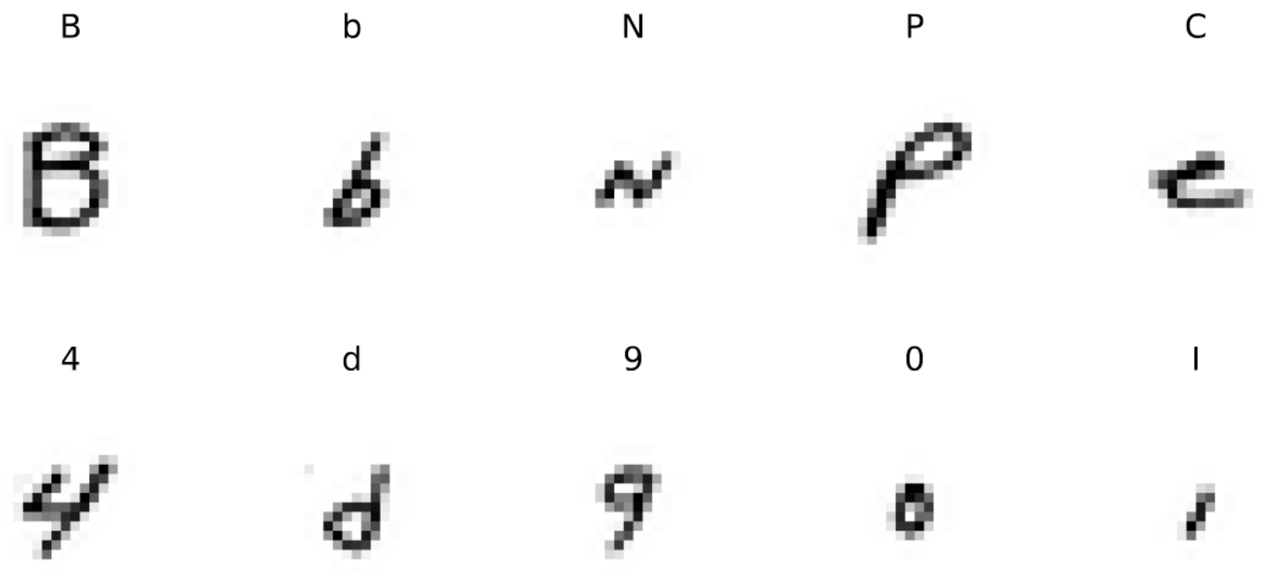
Experiments

- FEMNIST dataset
- $S(x,y) = \frac{1}{p_f(y|x)^{1/4}}$



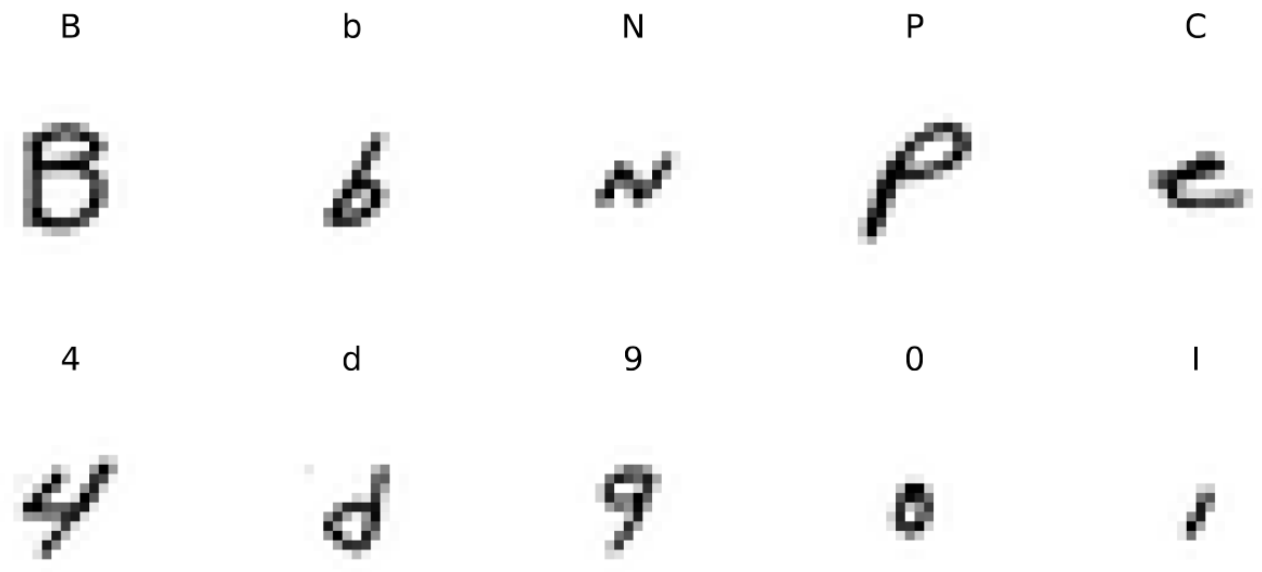
Experiments

- FEMNIST dataset
- $S(x, y) = \frac{1}{p_f(y|x)^{1/4}}$
- $\lambda = 1$



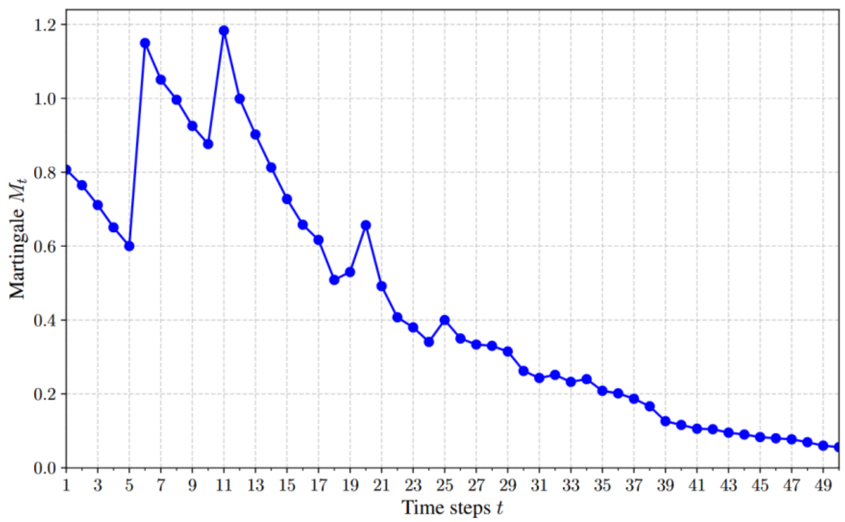
Experiments

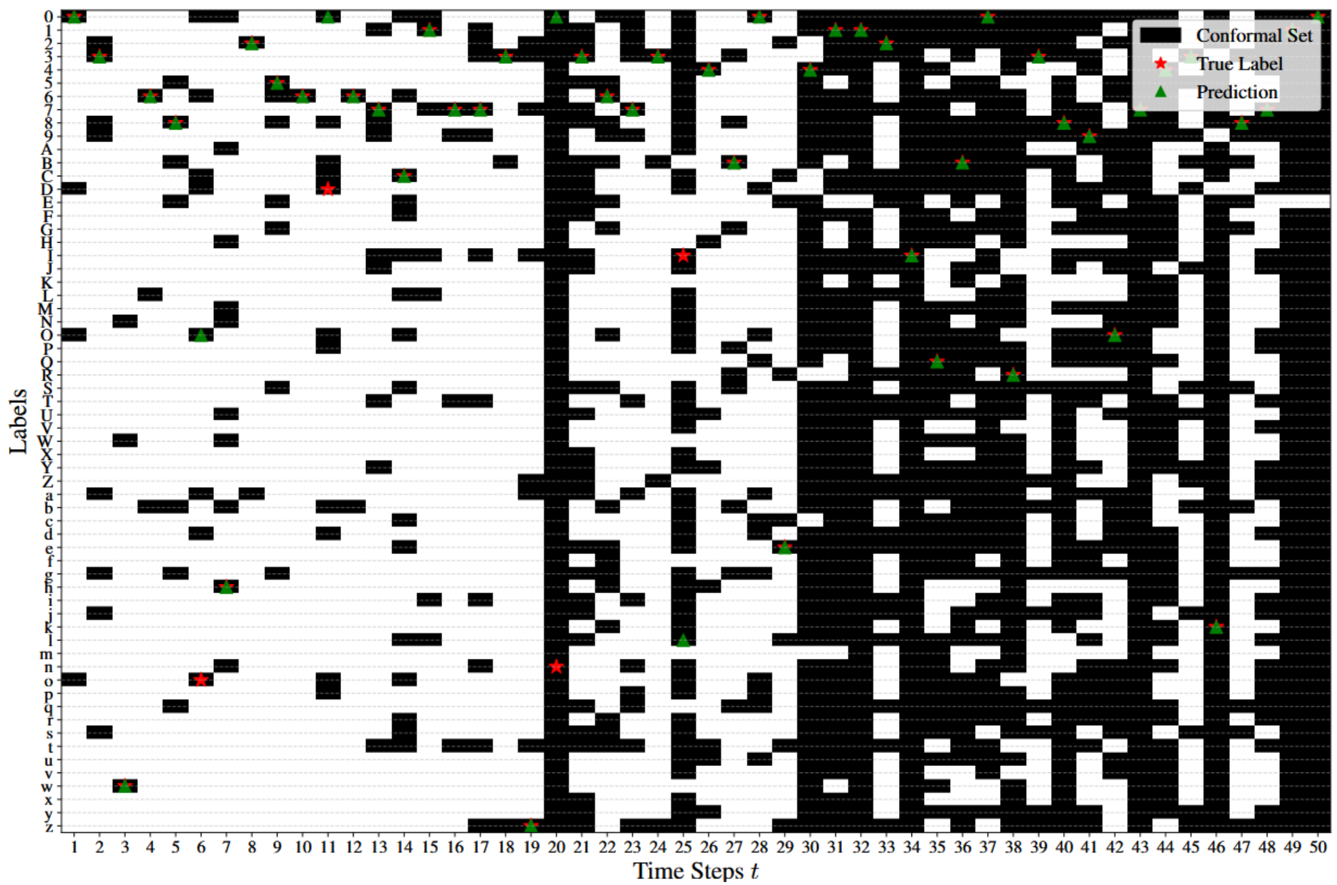
- FEMNIST dataset
- $S(x,y) = \frac{1}{p_f(y|x)^{1/4}}$
- $\lambda = 1$
- $\alpha = 0.15$



Experiments

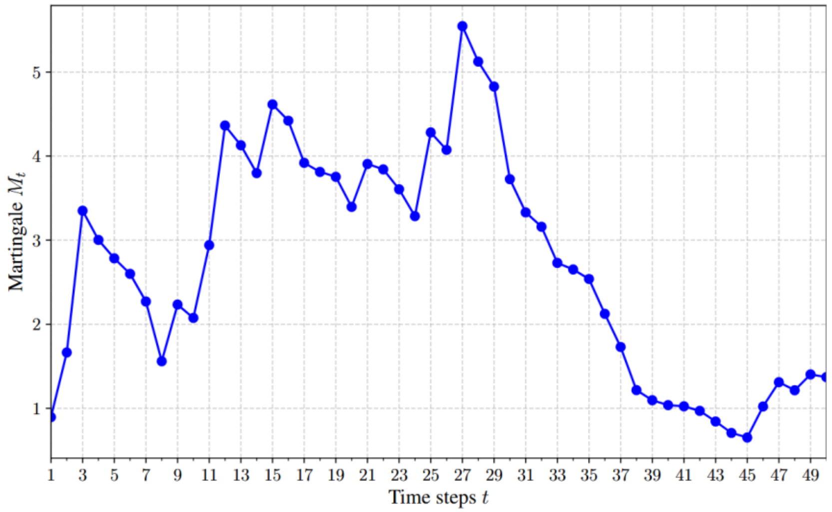
- FEMNIST dataset
- $S(x,y) = \frac{1}{p_f(y|x)^{1/4}}$
- $\lambda = 1$
- $\alpha = 0.15$

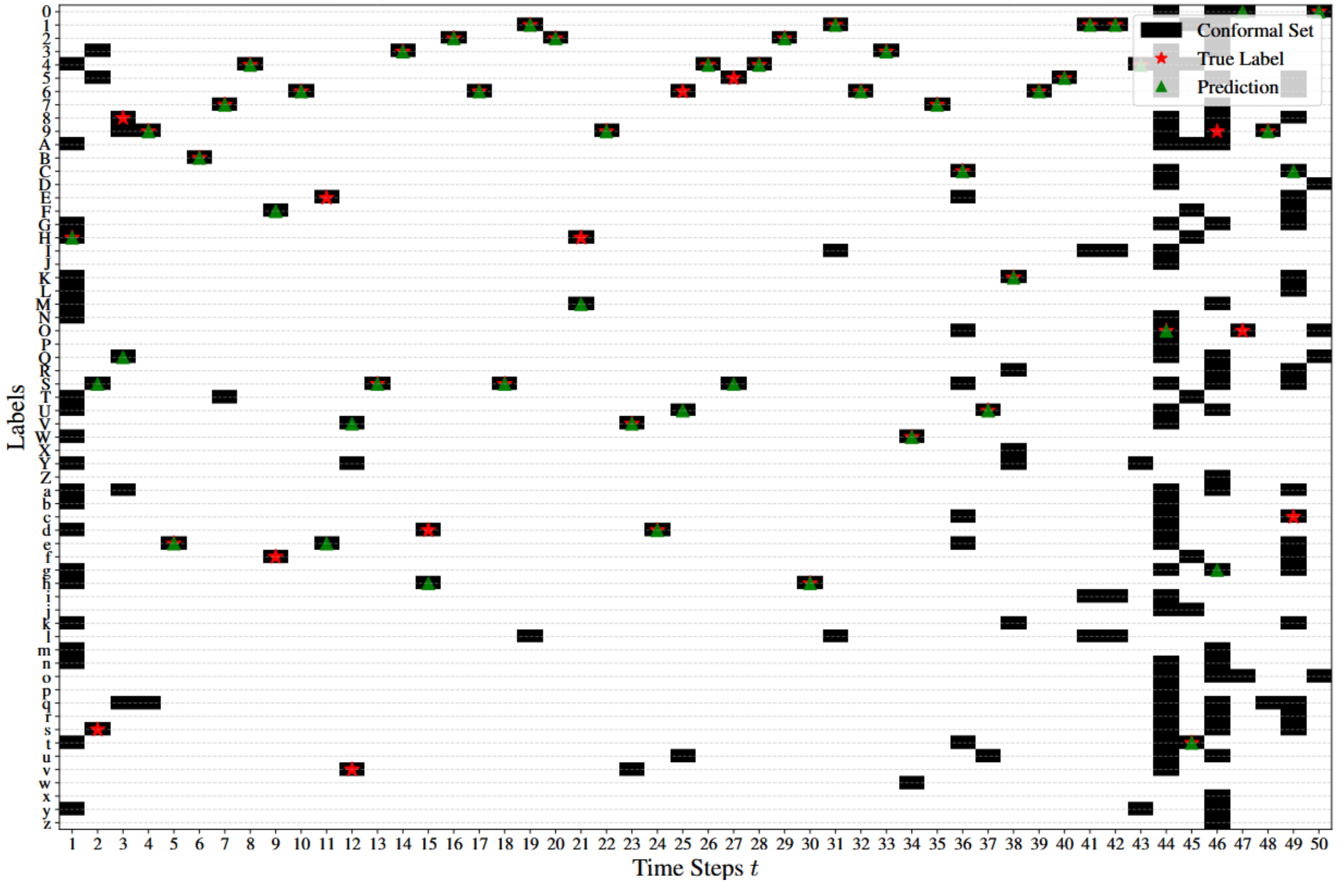




Experiments

- FEMNIST dataset
- $S(x,y) = \dfrac{1}{p_f(y|x)^{1/4}}$
- $\lambda = 1$
- $\alpha = 0.15$

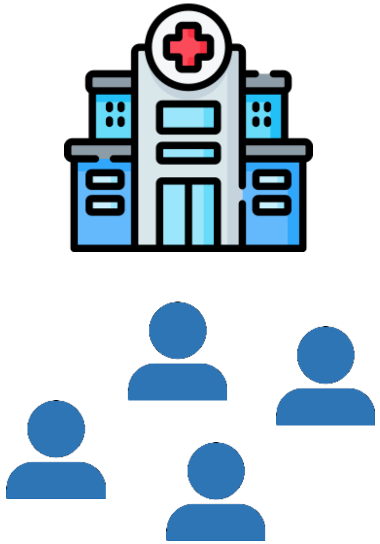




Today's Agenda

- Basics of Conformal Prediction
- Batch Anytime-valid Conformal Prediction
- Conformal Prediction with Adaptive Coverage
- Conformal Prediction under Ambiguous Ground Truth

Motivating Example

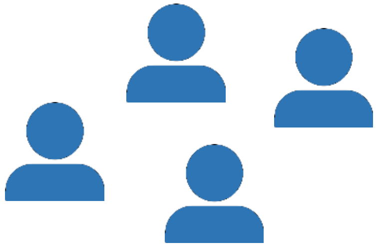


Calibration dataset

Test patient

{
Pneumonia
COVID-19
Asthma
COPD
Tuberculosis
Lung Cancer
Pulmonary Embolism
Bronchitis
Influenza
Sarcoidosis
}

Motivating Example



Calibration dataset

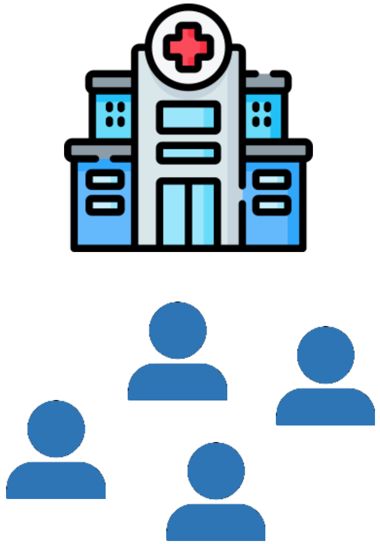
Test patient

$\alpha = 0.01$

$$\left\{ \begin{array}{c} \text{Pneumonia} \\ \text{COVID-19} \\ \text{Asthma} \\ \text{COPD} \\ \text{Tuberculosis} \\ \text{Lung Cancer} \\ \text{Pulmonary Embolism} \\ \text{Bronchitis} \\ \text{Influenza} \\ \text{Sarcoidosis} \end{array} \right\}$$

$$|C(X_{test})| = 8$$

Motivating Example



Calibration dataset

Test patient

$\alpha = 0.02$

$$\left\{ \begin{array}{c} \text{Pneumonia} \\ \text{COVID-19} \\ \text{Asthma} \\ \text{COPD} \\ \text{Tuberculosis} \\ \text{Lung Cancer} \\ \text{Pulmonary Embolism} \\ \text{Bronchitis} \\ \text{Influenza} \\ \text{Sarcoidosis} \end{array} \right\}$$

$|C(X_{test})| = 3$

Main Result

Post-hoc guarantees

Post-hoc p-variables [Wang & Ramdas 2022, Xu et al. 2024, Grünwald 2024, Ramdas & Wang 2024, Koning 2024]:

$$\sup_{\tilde{\alpha}>0} \mathbb{E}\left[\frac{\mathbb{P}(P \leq \tilde{\alpha} \mid \tilde{\alpha})}{\tilde{\alpha}}\right] \leq 1$$

P is a post-hoc p-variable if and only if $\mathbb{E}[1/P] \leq 1$.

Main Result

Post-hoc guarantees

Post-hoc p-variables [Wang & Ramdas 2022, Xu et al. 2024, Grünwald 2024, Ramdas & Wang 2024, Koning 2024]:

P

➢ **Conformal Prediction with Adaptive Coverage:**

$$\mathbb{E}\left[\frac{\mathbb{P}(Y_{test} \notin C(X_{test}) | \tilde{\alpha})}{\tilde{\alpha}}\right] \leq 1,$$

for any adaptive (possibly data-dependent) miscoverage $\tilde{\alpha} > 0$, where:

$$C(X_{test}) = \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < 1/\tilde{\alpha} \right\}.$$

Main Result
Post-hoc guarantees

Post-hoc p-variables [Wang & Ramdas 2022, Xu et al. 2024, Grünwald 2024, Ramdas & Wang 2024, Koning 2024]:

$$\left[\mathbb{P}(P \le \tilde{\alpha} \mid \tilde{\alpha})\right]$$

P

➢ **Conformal Prediction with Adaptive Coverage:**

$$\mathbb{E}\left[\frac{\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})}{\tilde{\alpha}}\right] \le 1,$$

for any adaptive (possibly data-dependent) miscoverage $\tilde{\alpha} > 0$, where:

$$C(X_{test}) = \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < 1/\tilde{\alpha} \right\}.$$

First-order Taylor approximation: $\mathbb{E}\left[\frac{\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})}{\tilde{\alpha}}\right] \approx \frac{\mathbb{E}[\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})]}{\mathbb{E}[\tilde{\alpha}]} = \frac{\mathbb{P}(Y_{test} \notin C(X_{test}))}{\mathbb{E}[\tilde{\alpha}]}$

$$\mathbb{P}(Y_{test} \in C(X_{test})) \ge 1 - \mathbb{E}[\tilde{\alpha}]$$

Main Result

Post-hoc guarantees

Post-hoc p-variables [Wang & Ramdas 2022, Xu et al. 2024, Grünwald 2024, Ramdas & Wang 2024, Koning 2024]:

$$\left[\mathbb{P}(P \leq \tilde{\alpha} \mid \tilde{\alpha})\right]$$

P

- **Conformal Prediction with Adaptive Coverage:**

$$\mathbb{E}\left[\frac{\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})}{\tilde{\alpha}}\right] \leq 1,$$

for any adaptive (possibly data-dependent) miscoverage $\tilde{\alpha} > 0$, where:

$$C(X_{test}) = \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < 1/\tilde{\alpha} \right\}.$$

First-order Taylor approximation: $\mathbb{E}\left[\frac{\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})}{\tilde{\alpha}}\right] \approx \frac{\mathbb{E}[\mathbb{P}(Y_{test} \notin C(X_{test}) \mid \tilde{\alpha})]}{\mathbb{E}[\tilde{\alpha}]} = \frac{\mathbb{P}(Y_{test} \notin C(X_{test}))}{\mathbb{E}[\tilde{\alpha}]}$

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \mathbb{E}[\tilde{\alpha}]$$

Can be estimated using the calibration set [Gauthier, Bach & Jordan 2025]

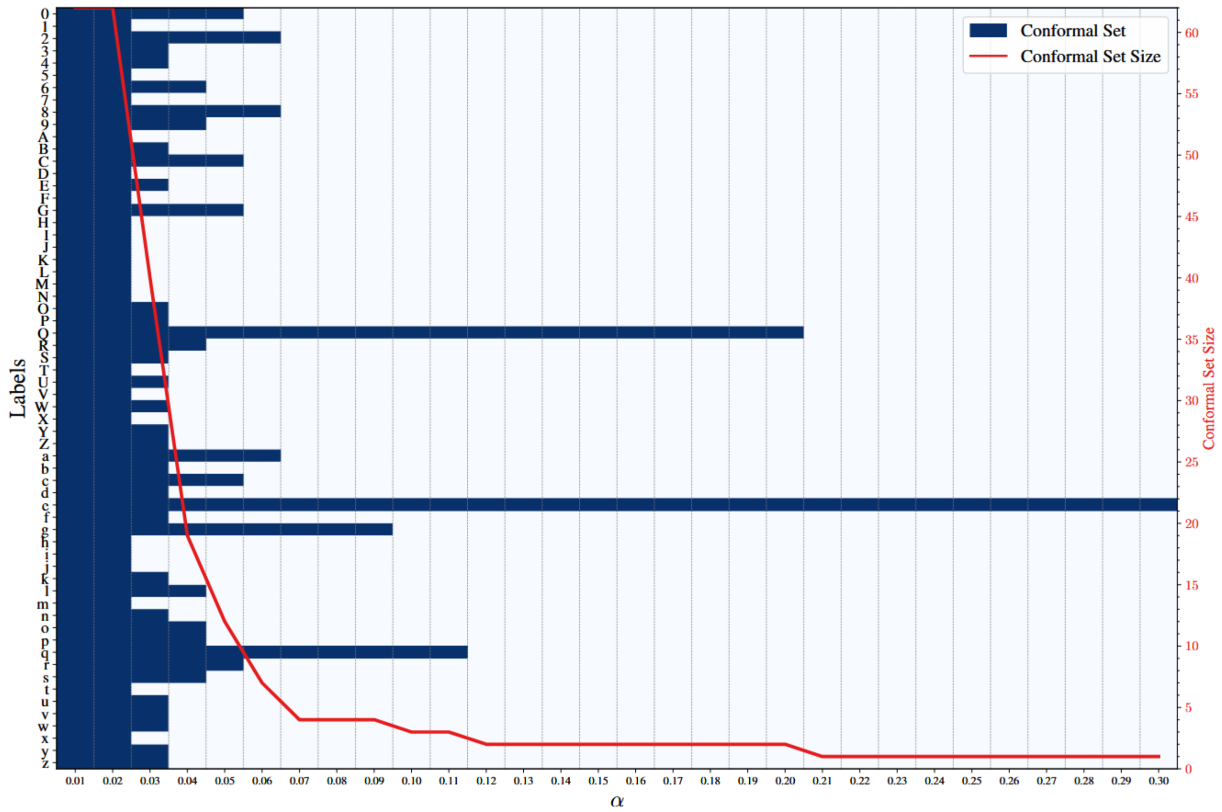
Experiments

$$\tilde{\alpha} = \inf \left\{ \alpha \in (0,1) : \# \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < \frac{1}{\alpha} \right\} \leq C(\{(X_i, Y_i)\}, X_{test}) \right\}$$

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \mathbb{E}[\tilde{\alpha}]$$

Experiments

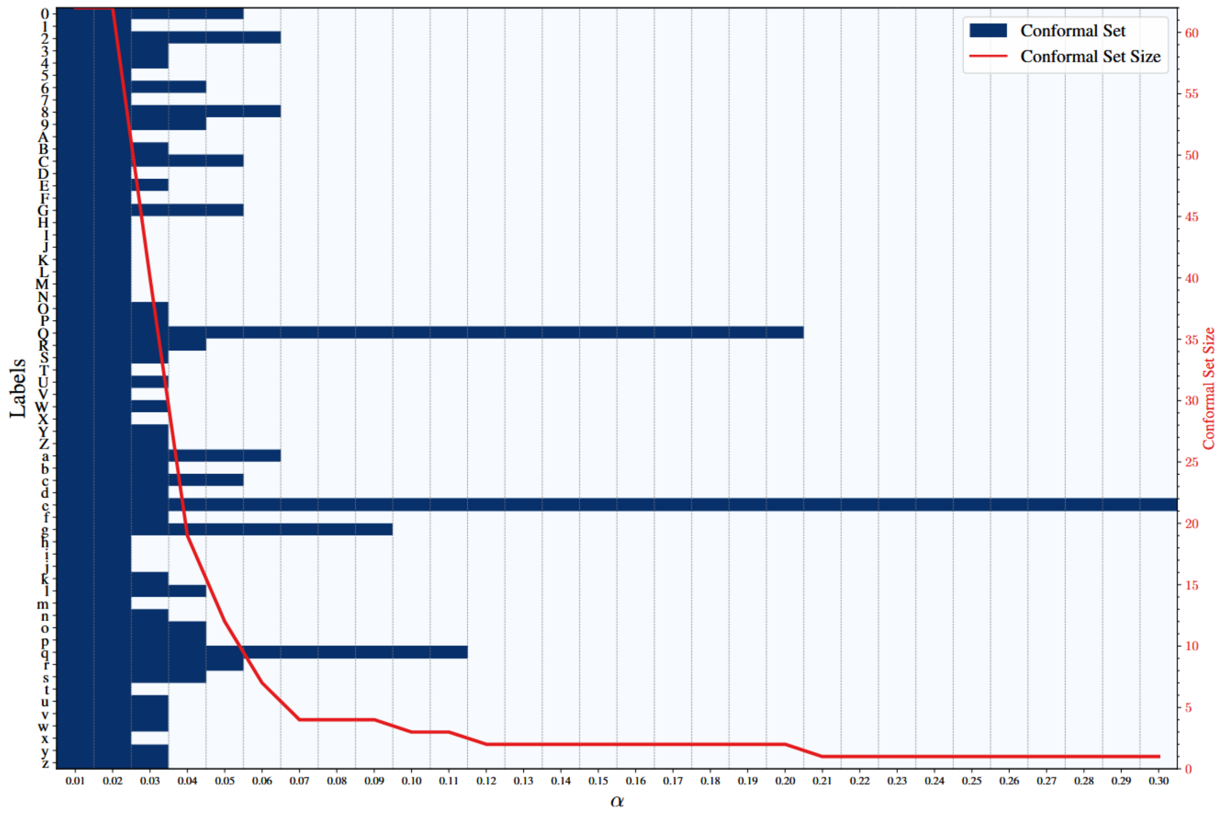
$$\tilde{\alpha} = \inf \left\{ \alpha \in (0,1) : \# \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < \frac{1}{\alpha} \right\} \leq C(\{(X_i, Y_i)\}, X_{test}) \right\}$$



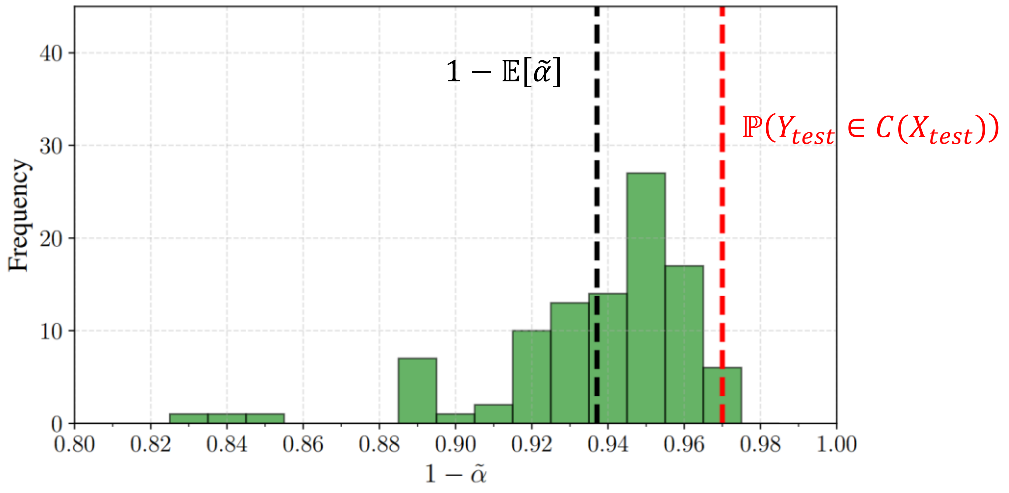
$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \mathbb{E}[\tilde{\alpha}]$$

Experiments

$$\tilde{\alpha} = \inf \left\{ \alpha \in (0,1) : \# \left\{ y : \frac{S(X_{test}, y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(X_{test}, y)\right)} < \frac{1}{\alpha} \right\} \leq C(\{(X_i, Y_i)\}, X_{test}) \right\}$$



$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \mathbb{E}[\tilde{\alpha}]$$



$C = 3$

Overview

- ❑ Basics of Conformal Prediction
- ❑ Batch Anytime-valid Conformal Prediction
- ❑ Conformal Prediction with Adaptive Coverage
- ❑ Conformal Prediction under Ambiguous Ground Truth

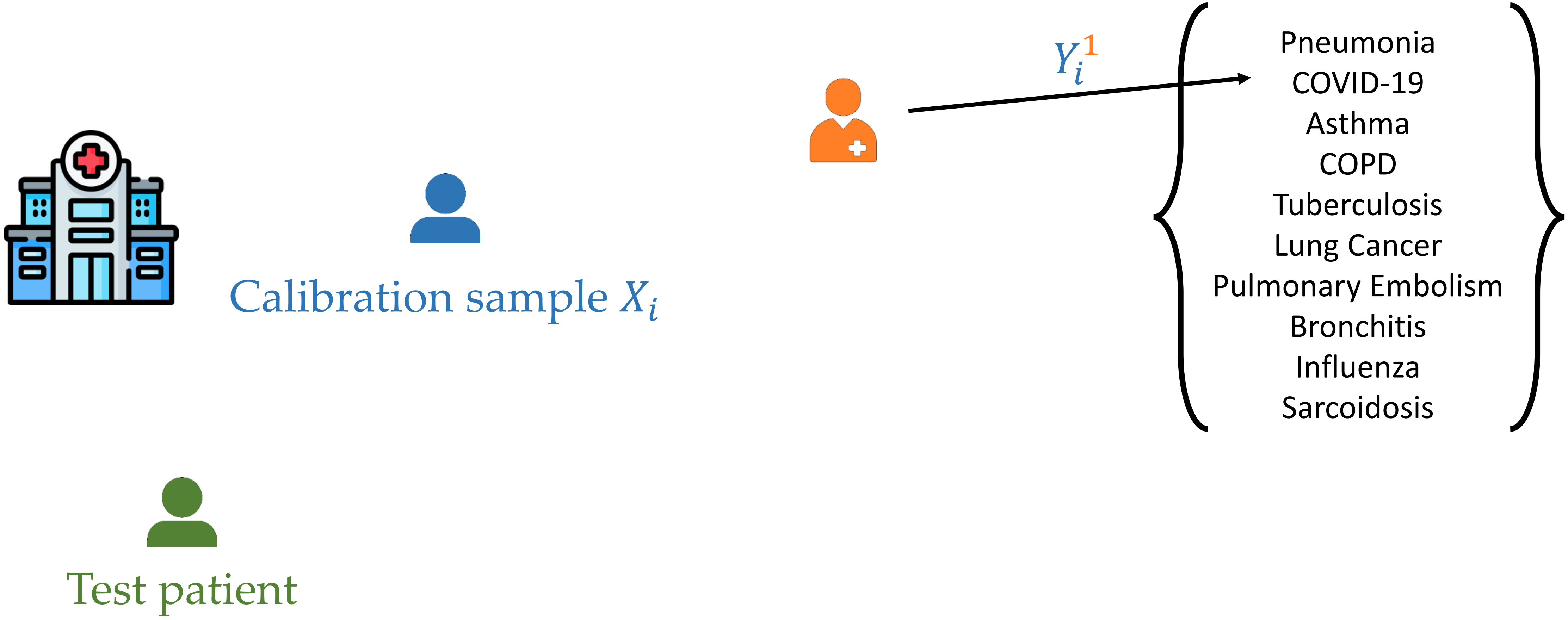
Motivating Example

Calibration sample X_i

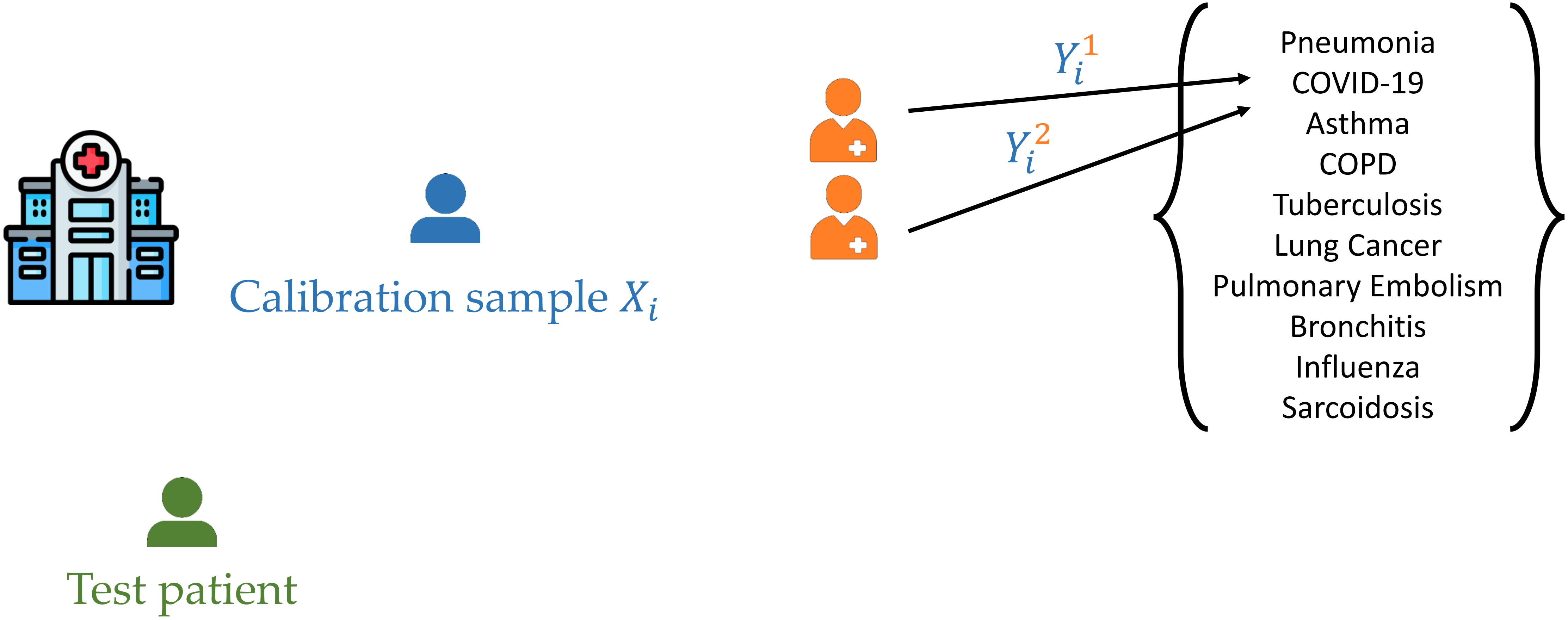
Test patient

Pneumonia
COVID-19
Asthma
COPD
Tuberculosis
Lung Cancer
Pulmonary Embolism
Bronchitis
Influenza
Sarcoidosis

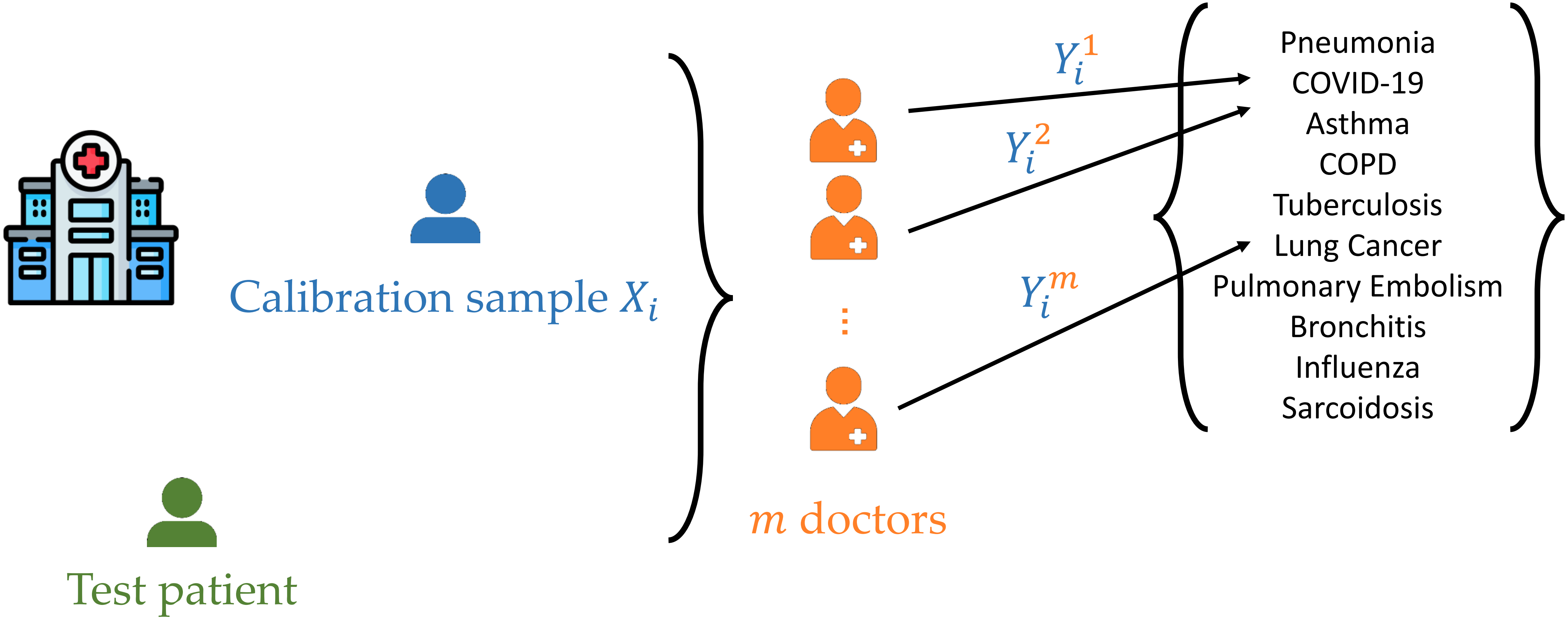
Motivating Example



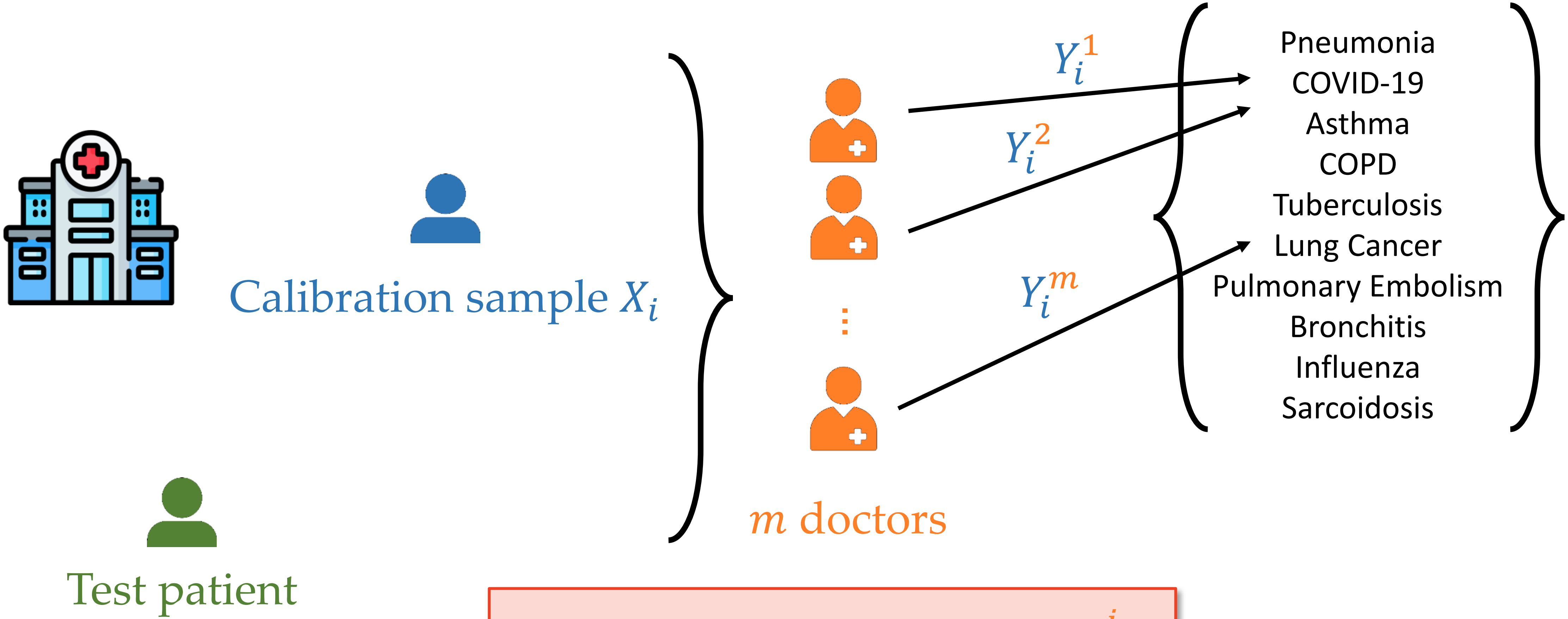
Motivating Example



Motivating Example



Motivating Example



➤ Calibration dataset: (X_i, Y_i^j)

Main Result

Average of e-values = e-value

$$E = \frac{1}{m}\sum_{j=1}^{m} E^{j}$$

$$E^{j} = \frac{S(X_{test}, Y_{test})}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i^{j}) + S(X_{test}, Y_{test})\right)}$$

Main Result

Average of e-values = e-value

$$E = \frac{1}{m}\sum_{j=1}^{m} E^j$$

$$E^j = \frac{S(X_{test}, Y_{test})}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i^j) + S(X_{test}, Y_{test})\right)}$$

➢ **Conformal Prediction under Ambiguous Ground Truth:**

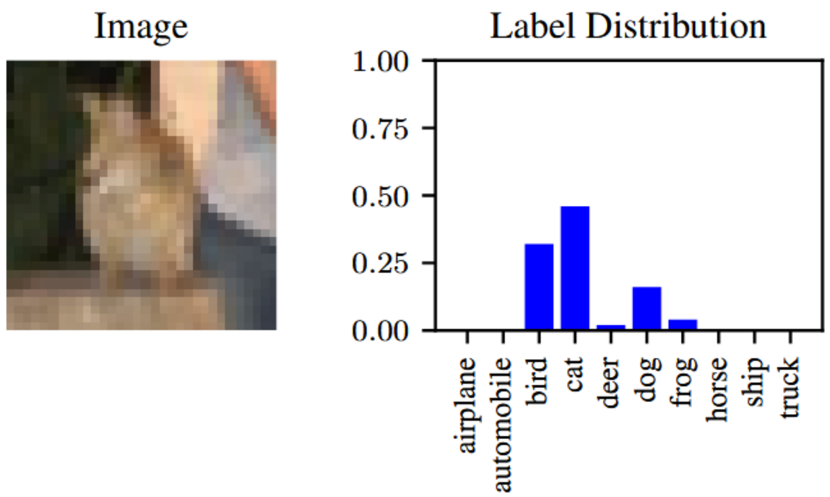
$$\mathbb{P}\left(Y_{test} \in C(X_{test})\right) \geq 1 - \alpha,$$

Where $C(X_{test}) = \left\{ y : \frac{1}{m}\sum_{j=1}^{m} \frac{S(X_{test}, Y_{test})}{\frac{1}{n+1}\sum_{i=1}^{n} S(X_i, Y_i^j) + S(X_{test}, Y_{test})} < 1/\alpha \right\}.$

Markov's *inequality*

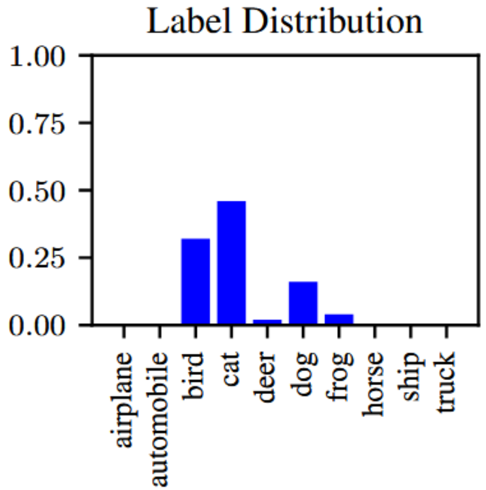
Experiments

- CIFAR-10H dataset (filtered)



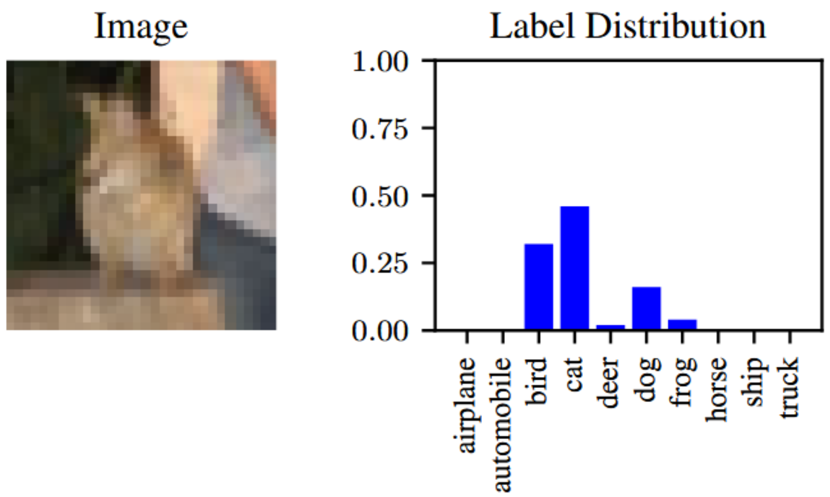
Experiments

- CIFAR-10H dataset (filtered)
- $S(x, y) = -\log p_f(y|x)$



Experiments

- CIFAR-10H dataset (filtered)
- $S(x,y) = -\log p_f(y|x)$



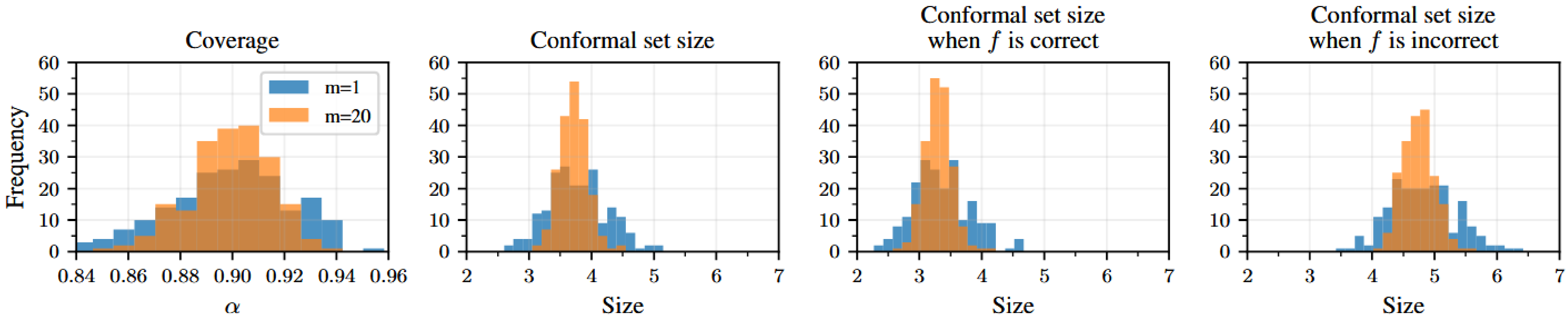


Figure 6: Comparison of coverage and conformal set sizes when using e-variables in Monte Carlo conformal prediction with $m = 1$ or $m = 20$ experts, with $\alpha = 0.3$, from Theorem 15.

Conclusion

- Explored **e-values for conformal prediction**, enabling more flexible inference
- Enables **online conformal methods** with anytime—valid guarantees
- Enables **data-dependent coverage guarantees**, allowing more adaptive and informative prediction sets tailored to individual test points
- Facilitates **easy aggregation of conformal prediction sets**, especially useful in cases of ambiguous ground truth
- Opens new avenues for conformal prediction:
  - Other possibilities for selecting data-dependent α…
- Open questions:
  - Choice of the score function in the soft-rank e-value?
  - Choice of the e-value?

Thank you! Questions?

Paper: